ORIGINAL ARTICLE

Journal Section

Spline-based nonparametric inference in general state-switching models

Roland Langrock[1] | **Timo Adam**[1] | **Vianey Leos-Barajas**[2] | **Sina Mews**[1] | **David L. Miller**[3] | **Yannis P. Papastamatiou**[4]

[1]Department of Business Administration and Economics, Bielefeld University, Bielefeld, 33615, Germany

[2]Department of Statistics, Iowa State University, Ames, IA 50011, USA

[3]Centre for Research into Ecological and Environmental Modelling, University of St Andrews, St Andrews, Fife, KY16 9LZ, UK

[4]School of Environment, Arts and Society, Florida International University, North Miami, FL 33181, USA

Correspondence
Roland Langrock, Department of Business Administration and Economics, Bielefeld University, Bielefeld, 33615, Germany
Email: roland.langrock@uni-bielefeld.de

Funding information
—

State-switching models combine immense flexibility with relative mathematical simplicity and computational tractability, and as a consequence have established themselves as general-purpose models for time series data. In this paper we provide an overview of ways to use penalised splines to allow for flexible nonparametric inference within state-switching models, and provide a critical discussion of the use of corresponding classes of models. The methods are illustrated using animal acceleration data and energy price data.

KEYWORDS
hidden Markov model; maximum penalised likelihood; Markov-switching regression; penalised splines

1 | INTRODUCTION

State-switching models assume that some observed process — e.g. a financial time series, a sequence of animal locations recorded with GPS, or blood samples repeatedly taken for a single individual — is

driven by an unobserved process that over time switches between different states, each of which implies a different probability model for the observations. Hidden Markov models (HMMs, Zucchini *et al.*, 2016) constitute the most prominent example of such a class of models, but there are several closely related models, e.g. Markov-switching regression models, general state-space models or Markov-modulated Poisson processes.

State-switching models provide natural frameworks for drawing comprehensive inference in diverse applied statistical problems arising in, *inter alia*, speech recognition (Juang and Rabiner, 1991), brain activity measurements (Langrock *et al.*, 2013), psychological learning experiments (Visser *et al.*, 2002), oceanic wave and wind modelling (Bulla *et al.*, 2012), records of volcanic eruptions (Bebbington, 2007), or animal abundance estimation subject to availability bias (Borchers *et al.*, 2013), to name but a few. Methodologically, a key asset of these classes of models is that an efficient recursive scheme, the so-called *forward algorithm*, can be applied to calculate the likelihood (Zucchini *et al.*, 2016), rendering the models convenient to work with despite their relatively complex dependence structure.

There are various methodological contributions in the area of state-switching models that focus on extensions of the basic model formulations, devising more complex dependence structures such as semi-Markov state processes (Guédon, 2003), coupled HMMs (Sherlock *et al.*, 2013), models with feedback from the observed to the state process (Zucchini *et al.*, 2008), or hierarchically structured state processes (Leos-Barajas *et al.*, 2017a). In addition, in the last decade, several papers have discussed inference for mixed HMMs for longitudinal data (Altman, 2007; Maruotti, 2011; Schliehe-Diecks *et al.*, 2012), where random effects are used to account for potential heterogeneity across multiple time series observed. Thus, considerable effort has gone into extending the *structure* of the basic model formulation.

However, we argue that a key component of state-switching models, namely the *probability model within states*, is often neglected in practical applications. Technological advancements have led to increasingly large data sets being collected, such that it is nowadays often possible to estimate the parameters of state-switching models with very high precision. For example, in animal movement modelling it is nowadays common to fit HMM-type models to hundreds of thousands of data points (see, e.g., Kock *et al.*, 2013, Morellet *et al.*, 2013, Lamb *et al.*, 2017). In these instances, even with state uncertainty there is sufficient information in the data to obtain an extremely detailed picture of state-dependent distributions of say distances travelled by an animal per given time unit. It is then often immediately clear that simple parametric distributions are inadequate to fully capture the shape of the empirical state-dependent distributions. Ignoring such a lack of fit can invalidate inference, for example on the number of states and on possible covariate effects on the state-switching dynamics (Langrock *et al.*, 2015b, Pohle *et al.*, 2017).

In a series of papers, we recently proposed the use of spline smoothing techniques for non-parametric inference within state-switching models (Langrock *et al.*, 2015a; Langrock *et al.*, 2015b; Langrock *et al.*, 2017, Adam *et al.*, 2017). The resulting classes of models combine two powerful tools, namely the forward algorithm for efficient likelihood evaluation, and penalised B-splines (i.e. P-splines; Eilers and Marx, 1996) for nonparametric inference, to allow for relatively straightforward and computationally tractable maximum penalised likelihood estimation within state-switching

models. The versatility of associated model formulations then opens up the way for various classes of models to be estimated nonparametrically. In this paper, we review these classes of models and associated spline-based estimation approaches, and discuss the benefits but also the caveats of working with nonparametric state-switching models. We distinguish state-switching density models (i.e. HMMs, Section 2) and state-switching regression models (i.e. Markov-switching regression models, Section 3).

2 | STATE-SWITCHING DENSITY MODELS

2.1 | Hidden Markov models

Hidden Markov models (HMMs) comprise two stochastic processes, only one of which is observed. The observed process is a time series $\{Y_t\}_{t=1,\ldots,T}$, the observations of which can be either discrete or continuous, and also multivariate. Here we focus on the case of univariate continuous observations. In an HMM it is assumed that each observation is generated by one of N component distributions, as selected by the state of the unobserved state process $\{S_t\}_{t=1,\ldots,T}$, and that conditional on the states, the observations are independent of each other. In its most basic form, the state process is assumed to be an N-state Markov chain, and typically exhibits persistence in the different states, thereby inducing serial correlation in the observed time series. The dependence structure of such a basic HMM is illustrated in Figure 1.

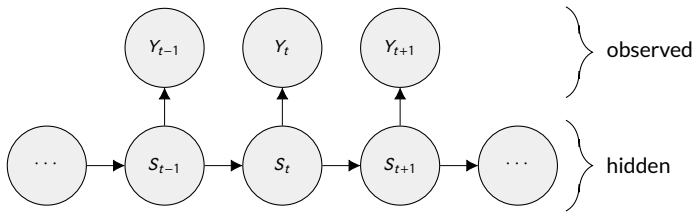

FIGURE 1 Dependence structure of the most basic univariate hidden Markov model.

Due to the simple dependence structure, in the time-homogeneous case such a basic HMM is fully specified by the transition probability matrix (t.p.m.), $\boldsymbol{\Gamma} = (\gamma_{ij})$, with $\gamma_{ij} = \Pr(S_{t+1} = j \,|\, S_t = i)$, the initial state distribution, $\boldsymbol{\delta} = \big(\Pr(S_1 = 1), \ldots, \Pr(S_1 = N)\big)$ (often taken to be the stationary distribution implied by $\boldsymbol{\Gamma}$), and the densities $f_1(y), \ldots, f_N(y)$ of the state-dependent distributions (using the shorthand notation $f_i(y) = f(y \,|\, S_t = i)$, $i = 1, \ldots, N$). Thus, to formulate an HMM, one needs to choose the number of states, N, and the class of probability distributions from which the $f_i(y)$ are taken. The former is essentially a model selection problem (albeit a hard one, see Pohle *et al.*, 2017). The choice of the distributional family is usually dictated by the data at hand. For example, for log-returns on shares, which are real numbers, it is common to assume either normal or, if more flexibility to accommodate heavy tails is desired, Student-t state-dependent distributions (Bulla and Bulla, 2006). In animal movement modelling, key quantities commonly considered are the step

lengths an animal performs between consecutive locations at which it is observed. Step lengths are, by nature, positive real numbers, and thus gamma or Weibull state-dependent distributions are commonly assumed (Michelot *et al.*, 2016).

2.2 | Motivation for nonparametric inference within HMMs

While conceptually straightforward, choosing an adequate parametric family for the state-dependent distributions in practice is by no means a trivial task. As in simple univariate density estimation, it is often the case that the actual shape of the state-dependent distributions is complex, e.g. exhibiting heavy tails, skewness, or even multimodality. This is exacerbated by the fact that there is no way of visualising the empirical distributions within a state *a priori* (here: before a model has been fitted), since it is unknown which observations are associated with which underlying state. In practice, it is thus often very difficult to choose an adequate parametric family.

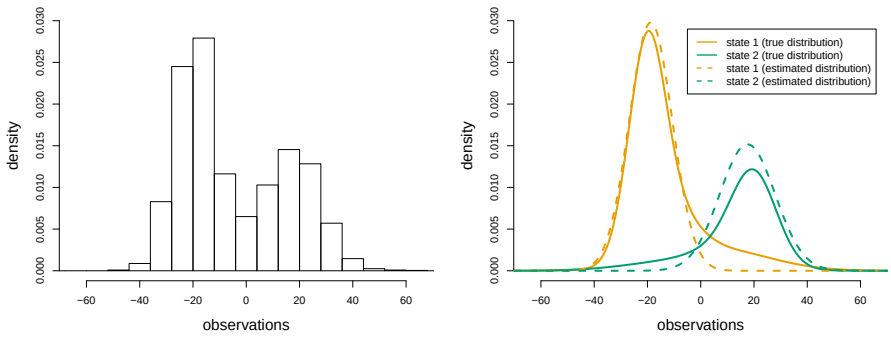

FIGURE 2 Left plot: histogram of 3000 observations generated from a 2-state HMM with skewed state-dependent distributions. Right plot: true state-dependent distributions used to generate the data (solid lines) and estimated normal state-dependent distributions (dashed lines) — both the true and the estimated distributions here are weighted with the stationary probabilities of the Markov chain occupying the different states (true and estimated, respectively).

We illustrate this point in Figure 2. The left plot displays a histogram of 3000 observations that were simulated from a 2-state HMM. Based on this estimator of the marginal distribution of the observations, it is tempting to conclude that two distinct normal distributions, with means roughly at −20 and 20, respectively, may have generated the data. However, the actual state-dependent distributions used to generate the data were heavily right skewed (state 1) and left skewed (state 2), respectively. The right plot in Figure 2 displays the consequences of fitting a misspecified 2-state HMM with normal state-dependent distributions to these data. It may be argued here that the deviation of the fitted from the actual state-dependent distributions is only relatively minor. However, even such a minor mismatch can have various undesirable consequences, including:

- a poor predictive performance (which would be problematic for example in financial risk management applications);
- frequent misclassification of observations particularly in the areas of overlap between the state-dependent distributions (which could be problematic if interest primarily lies in decoding the hidden states, for example in recognition tasks);
- invalid inference on state-switching dynamics (e.g. regarding potential covariate influence on state transition probabilities);
- invalid inference on the number of states.

The last point above was discussed in detail in Langrock *et al.* (2015b) and Pohle *et al.* (2017). In short, if a misspecified model such as the 2-state HMM with normal distributions is fitted, then model selection criteria will point to models with more states than actually present in the data, with the additional states being included to "mop up" the structure that is not being accounted for, in this case the heavy tails of the state-dependent distributions. While the example above is artificial, this is indeed a major problem in practical applications of HMMs, where simple parametric distributions are rarely sufficiently flexible to capture the key features of empirical distributions within states, such that conducting model selection on N will inevitably lead to overly complex state architectures.

An obvious way to overcome the insufficient flexibility of the 2-state normal HMM above would be to use mixtures of say normal distributions within states (Volant *et al.*, 2015; Holzmann and Schwaiger, 2015; Leos-Barajas *et al.*, 2017b). The number of mixture components to be used for each state can then be determined using model selection criteria or testing. Alternatively, a very large number of mixture components, say 30, can be used, then including a suitable penalty term in the likelihood as to avoid overfitting. In such HMMs, the state-dependent distributions are constructed as linear combinations of much simpler densities. Not only the weight in the mixture, but also the location and scale of these densities is estimated from the data. While feasible for very small numbers of components, this quickly leads to severe numerical instability if many component distributions need to be considered, thus rendering this approach less useful for distributions with particularly complex shapes. Alternatively, one could specify a large number of *fixed* basis densities, then estimating only the weights of these in the linear combination that ultimately yields the state-dependent distribution. This strategy is explained in more detail in the subsequent section.

2.3 | B-spline-based nonparametric model formulation

There are different ways to specify a set of basis densities from which a (state-dependent) distribution can be constructed via linear combination. Langrock *et al.* (2015b) suggested using B-splines which, in particular, form an efficient and convenient basis (de Boor, 1978; Eilers and Marx, 1996). In state i, $i = 1, \ldots, N$, the state-dependent distribution is then formulated as follows:

$$f_i(y) = \sum_{k=-K}^{K} \omega_{k,i} \phi_k(y), \tag{1}$$

with a set of regularly spaced B-splines $\phi_{-K}, \ldots, \phi_K$, standardised such that they integrate to one. In order to ensure that $f_i(y)$ is a probability density function, the coefficients to be estimated, $\omega_{-K,i}, \ldots, \omega_{K,i}$, are reparameterised using the multinomial logit link:

$$\omega_{k,i} = \frac{\exp(\beta_{k,i})}{\sum_{j=-K}^{K} \exp(\beta_{j,i})},$$

such that the (unconstrained) coefficients $\beta_{k,i}$ are estimated. Using this transformation, the resulting $\omega_{k,i}$ (and hence also $f_i(y)$) are non-negative and sum to one (such that $\int f_i(y)dy = 1$ due to the standardisation of the B-spline basis functions). We set $\beta_{i,0} = 0$ for identifiability. Cubic B-splines are twice continuously differentiable and hence yield visually smooth density estimates, such that they constitute a suitable default.

When estimated nonparametrically, the marginal distribution of the observations could in fact already be captured using just one distribution as formulated in (1), i.e. a single-state model. However, such a model would not capture any serial correlation. If there is correlation in the time series — typically such that there is persistence in the states — then nonparametric HMMs with multiple states are identifiable. Mathematically, identifiability in nonparametric HMMs holds if the t.p.m. has full rank and the state-dependent distributions are distinct (Alexandrovich *et al.*, 2016). In practice, these conditions will usually be satisfied.

Using the artificial example discussed above, Figure 3 illustrates how the state-dependent distributions are constructed as linear combinations of weighted B-spline basis densities. Fitting this model to the simulated data was achieved by numerically maximising the (penalised) log-likelihood with respect to the coefficients $\omega_{i,k}$, $i = 1, \ldots, N$, $k = -K, \ldots, K$, as well as the t.p.m. $\boldsymbol{\Gamma}$. A relatively large number of basis elements, 51 (hence $K = 25$), was chosen to obtain virtually unlimited flexibility for capturing complex distributional shapes. A penalty term was added to the log-likelihood to control the wiggliness of the fitted distribution (and thus avoid overfitting). More details on this penalised likelihood approach are provided in the subsequent section.

2.4 | Inference

2.4.1 | Direct numerical maximum penalised likelihood

The most straightforward way to fitting an HMM, either parametric or nonparametric, to data is via numerical maximisation of the likelihood with respect to the parameters. In order to evaluate the likelihood, a recursive scheme called the *forward algorithm* is used. For this, we consider the forward variables at time t,

$$\boldsymbol{\alpha}_t = \big(\alpha_t(1), \ldots, \alpha_t(N)\big), \quad \text{where} \quad \alpha_t(i) = f(y_1, \ldots, y_t, s_t = i), \quad i = 1, \ldots, N.$$

The variable $\boldsymbol{\alpha}_t$ contains information on the likelihood of all observations up to time t (since $f(y_1, \ldots, y_t) = \sum_{i=1}^{N} \alpha_t(i)$), while retaining information on the probabilities of the process being in the different states (since $\Pr(S_t = i | y_1, \ldots, y_t) = \alpha_t(i) / \sum_{i=1}^{N} \alpha_t(i)$). Crucially, the forward variables can be calcu-

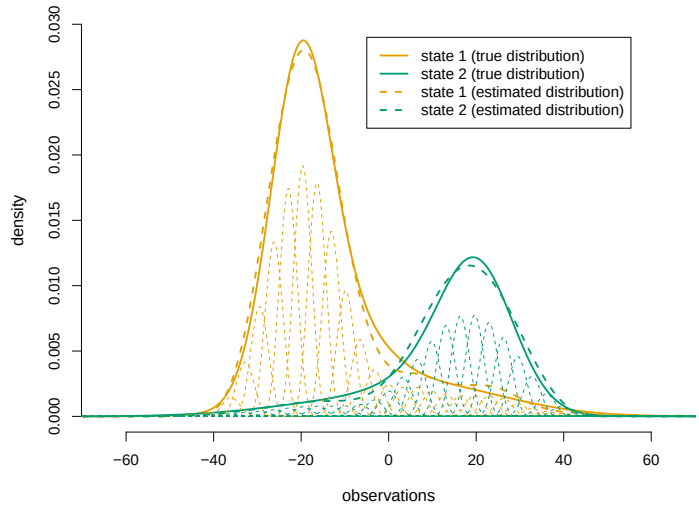

FIGURE 3 True state-dependent distributions (solid lines) and state-dependent distributions estimated using linear combinations of weighted B-splines (dashed lines). Both the true and the estimated distributions here are weighted with the stationary probabilities of the Markov chain occupying the different states (true and estimated, respectively). Displayed below the densities are the contributions of the individual B-spline basis functions to the density estimators.

lated recursively, beginning at time $t = 1$, then traversing along the time series and updating $\boldsymbol{\alpha}_t$ along the way, as follows:

$$\boldsymbol{\alpha}_1 = \boldsymbol{\delta}\mathbf{P}(y_1), \qquad \boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(y_t), \tag{2}$$

where $\mathbf{P}(y_t) = \text{diag}\big(f_1(y_t), \ldots, f_N(y_t)\big)$. The likelihood is then obtained as

$$\mathcal{L} = f(y_1, \ldots, y_T) = \sum_{i=1}^{N} \alpha_T(i) = \boldsymbol{\alpha}_T \mathbf{1}^\top, \tag{3}$$

where $\mathbf{1} \in \mathbb{R}^N$ is a row vector of ones. Notably, the computational effort involved in evaluating $\mathcal{L}$ is only linear in T, the number of observations, which opens up the way for numerical maximum likelihood even for long time series. Technical issues arising in the numerical maximisation of $\mathcal{L}$, such as parameter constraints, numerical underflow, and local maxima, are discussed in detail in Chapter 3 of Zucchini *et al.* (2016). In particular, a scaling strategy can be applied to calculate the log-likelihood, which is used in the penalised spline estimation approach presented below. To guard

against missing the global maximum of the penalised log-likelihood, the main strategy in practice is to run the algorithm many times using random initial points.

When B-splines are used to construct the densities of the state-dependent distributions, via linear combination as in (1), then the diagonal entries of the matrices $\mathbf{P}(y_t)$, $t = 1, \ldots, T$, can conveniently be calculated as

$$\begin{pmatrix} f_1(y_1) & \ldots & f_N(y_1) \\ \vdots & & \vdots \\ f_1(y_T) & \ldots & f_N(y_T) \end{pmatrix} = \mathbf{B}\boldsymbol{\Omega},$$

where

$$\mathbf{B} = \begin{pmatrix} \phi_{-K}(y_1) & \ldots & \phi_K(y_1) \\ \vdots & & \vdots \\ \phi_{-K}(y_T) & \ldots & \phi_K(y_T) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Omega} = \begin{pmatrix} \omega_{-K,1} & \ldots & \omega_{-K,N} \\ \vdots & & \vdots \\ \omega_{K,1} & \ldots & \omega_{K,N} \end{pmatrix}.$$

The entries $\omega_{k,i}$ are parameters to be estimated, while the design matrix $\mathbf{B}$ is fixed. To avoid overfitting, we penalise the log-likelihood by adding a wiggliness penalty as in Eilers and Marx (1996):

$$l_{\text{pen}} = \log \mathcal{L} - \sum_{i=1}^{N} \frac{\lambda_i}{2} \sum_{-K+m}^{K} (\Delta^m \omega_{k,i})^2, \tag{4}$$

where $\Delta\omega_k = \omega_k - \omega_{k-1}$ and $\Delta^m\omega_k = \Delta(\Delta^{m-1}\omega_k)$. The $\lambda_1, \ldots, \lambda_N$ are state-specific smoothing parameters, which control the influence of the penalty for each state. Letting $m = 2$, we obtain second-order differences between adjacent B-spline coefficients. This provides an approximation to the integrated squared second derivatives (the minimiser which we might definitionally consider to be "smoothest"; Green and Silverman, 1994). However, third-order differences are also theoretically appealing (Eilers and Marx, 1996).

Estimation of the smoothing parameters for nonparametric HMMs is underdeveloped at this point. Current approaches include: cross-validation, model selection criteria that take into account the penalisation (thus estimating the effective degrees of freedom, rather than simply counting parameters, to measure model complexity), or subjective selection based on visual inspection of fitted models. Regarding the former two (formal) methods, we found that while they mostly produce reasonable values for the λ_i, they are nevertheless somewhat unstable and sometimes fail completely (Langrock *et al.*, 2015b; Langrock *et al.*, 2017). Overall, smoothing parameter selection remains a challenging task in these model classes. In the subsequent section, we outline how this situation can potentially be improved upon by using the expectation-maximisation (EM) algorithm, rather than direct numerical maximisation, for finding the maximum penalised likelihood estimate.

2.4.2 | Estimation using the EM algorithm

An alternative way to fitting an HMM to data, which also arrives at the maximum likelihood estimate, is via the EM algorithm (Baum *et al.*, 1970; Dempster *et al.*, 1977; Welch, 2003). The basic idea of the EM algorithm is to iteratively impute missing data or latent variables (within HMMs: the states) conditional on the observed data and a given set of model parameters (which is referred to as the E step), then update the parameters using the current guess of the missing data (which is referred to as the M step), and so forth until convergence. Implementation of the EM algorithm for HMMs is technically more involved than direct numerical likelihood maximisation (MacDonald, 2014), such that the latter approach will usually be preferable. However, the EM algorithm can be more robust in terms of finding the global maximum of the likelihood (Bulla and Berzel, 2008). In addition, it is sometimes advantageous to be able to maximise the parameters for given latent states (as done in the M step), which we discuss below for the case of nonparametric HMM formulations using splines.

Assuming the state sequence $s_1, \dots, s_T$ of an HMM to be observed, and defining the quantities $u_i(t) = 1_{\{s_t=i\}}$ and $v_{ij}(t) = 1_{\{s_{t-1}=i, s_t=j\}}$ for $i, j = 1, \dots N$, $t = 1, \dots, T$, the complete-data log-likelihood (CDLL), i.e. the joint log-likelihood of the observations *and* the states, can be written as

$$\begin{aligned} l_c = \log \mathcal{L}_c &= \log\left(\delta_{s_1} \prod_{t=2}^{T} \gamma_{s_{t-1},s_t} \prod_{t=1}^{T} f_{s_t}(y_t)\right) \\ &= \log(\delta_{s_1}) + \sum_{t=2}^{T} \log(\gamma_{s_{t-1},s_t}) + \sum_{t=1}^{T} \log(f_{s_t}(y_t)) \\ &= \sum_{i=1}^{N} u_i(1) \log(\delta_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T} v_{ij}(t) \log(\gamma_{ij}) + \sum_{i=1}^{N}\sum_{t=1}^{T} u_i(t) \log\big(f_i(y_t)\big), \end{aligned}$$

with the associated complete-data *penalised* log-likelihood (CDPLL)

$$l_{c,\text{pen}} = l_c - \sum_{i=1}^{N} \frac{\lambda_i}{2} \sum_{-K+m}^{K} (\Delta^m \beta_{k,i})^2.$$

Note that the different parameters (i.e., the δ_i, the γ_{ij}, and the $\beta_{k,i}$) appear in distinct summands, which greatly simplifies the maximisation in the M step. Unlike in (4), here we penalise wiggliness by considering m-th order differences of the unconstrained $\beta_{k,i}$ parameters (rather than considering the $\omega_{k,i}$), in order to be able to follow the approach developed in Schellhase and Kauermann (2012), as detailed below.

However, the state sequence is of course not actually observed, and the main idea of the EM algorithm is to alternate between guessing the states (given the parameters) and updating the parameter values based on the CDPLL (given the states). More specifically, expressing the state sequence in terms of the indicator variables $u_i(t)$ and $v_{ij}(t)$ defined above, we calculate the conditional expectations of these given the current parameter values and the data, which is straightforward using the forward and backward variables. The forward variables are defined as in the previous section, and

the backward variables at time t as

$$\boldsymbol{\beta}_t = \big(\beta_t(1), \ldots, \beta_t(N)\big), \quad \text{where} \quad \beta_t(j) = f(y_{t+1}, \ldots, y_T \,|\, S_t = j), \quad j = 1, \ldots, N.$$

Analogously as in case of the forward variables, the backward variables can be calculated recursively using the *backward algorithm*, beginning at time $t = T$, then traversing backwards through time:

$$\boldsymbol{\beta}_T = \mathbf{1}, \qquad \boldsymbol{\beta}_t^\top = \boldsymbol{\Gamma}\mathbf{P}(y_{t+1})\boldsymbol{\beta}_{t+1}^\top.$$

The E step involves calculating the conditional expectations of the $u_i(t)$ and of the $v_{ij}(t)$, respectively, given the data and the current parameter estimates, as detailed in the following.

1. It follows from the definition of the forward and backward probabilities that

$$\hat{u}_i(t) = \Pr(S_t = i \mid y_1, \ldots, y_T) = \frac{\boldsymbol{\alpha}_t(i)\boldsymbol{\beta}_t(i)}{\boldsymbol{\alpha}_T\mathbf{1}^\top}$$

   for $t = 1, \ldots, T$, $i = 1, \ldots, N$.
2. It follows from the definition of the forward, backward and state transition probabilities that

$$\hat{v}_{ij}(t) = \Pr(S_{t-1} = i, S_t = j \mid y_1, \ldots, y_T) = \frac{\boldsymbol{\alpha}_{t-1}(i)\gamma_{ij} f_j(y_t)\boldsymbol{\beta}_t(j)}{\boldsymbol{\alpha}_T\mathbf{1}^\top}$$

   for $t = 2, \ldots, T$, $i, j = 1, \ldots, N$.

The M step involves the maximisation of the CDPLL — where the $u_i(t)$ and $v_{ij}(t)$ are replaced by their current estimates obtained in the previous E step — with respect to the model parameters:

1. Maximising the CDPLL with respect to δ_i yields the closed-form solution

$$\hat{\delta}_i = \hat{u}_i(1)$$

   for $i = 1, \ldots, N$.
2. Maximising the CDPLL with respect to γ_{ij} yields the closed-form solution

$$\hat{\gamma}_{ij} = \frac{\sum_{t=2}^{T} \hat{v}_{ij}(t)}{\sum_{k=1}^{N}\sum_{t=2}^{T} \hat{v}_{ik}(t)}$$

   for $i, j = 1, \ldots, N$.
3. Maximising the CDPLL with respect to the $\beta_{k,i}$ is slightly more involved, as they appear in both the third and the fourth summand, the latter of which also depends on the smoothing parameters λ_i. Similar to direct numerical maximum penalised likelihood, a natural approach is to estimate the $\beta_{k,i}$ for some fixed λ_i and select among different values using generalised cross-validation

or information criteria. However, although theoretically straightforward, these approaches typically require a grid search and are therefore computationally intensive. To overcome this drawback, we estimate the λ_i *within* each M step, using a (linear) mixed model representation (Schellhase and Kauermann, 2012). Now the penalty matrices can be considered to be prior precision matrices for the $\boldsymbol{\beta}_{k,i}$ (which are considered as random effects). The model can then be estimated using restricted maximum likelihood estimation (REML; Kauermann, 2005; Wood, 2011). Following Schellhase and Kauermann (2012), an estimating equation for the λ_i can be obtained from differentiating the linear mixed model log-likelihood with respect to λ_i, yielding the equation

$$\hat{\lambda}_i^{-1} = \frac{\hat{\boldsymbol{\beta}}_i^\top \mathbf{D}_m \hat{\boldsymbol{\beta}}_i}{\mathrm{df}(\hat{\lambda}_i) - (m-1)}, \tag{5}$$

where the effective degrees of freedom, $\mathrm{df}(\hat{\lambda}_i)$, can be approximated by $\mathrm{df}(\hat{\lambda}_i) = \mathrm{tr}(J_p^{-1}(\hat{\boldsymbol{\beta}}_i; \lambda_i = \hat{\lambda}_i)\, J_p(\hat{\boldsymbol{\beta}}_i; \lambda_i = 0))$ with $J_p(\boldsymbol{\beta}_i; \lambda_i)$ denoting the Hessian matrix of the CDPLL with respect to $\boldsymbol{\beta}_i$ for some fixed λ_i. Note that both sides of (5) depend on $\hat{\lambda}_i$, such that the solution needs to be found iteratively. Using the resulting estimate $\hat{\lambda}_i$, the relevant term in the CDPLL can then be maximised with respect to the $\boldsymbol{\beta}_{k,i}$ using some Newton-Raphson-type optimisation routine.

The EM algorithm alternates between the E and the M step until some convergence threshold is satisfied.

2.5 | Case study: oceanic whitetip shark acceleration data

HMMs are prominently used as tools for modelling animal movement data because of their ability to connect observed movement metrics to underlying states, where these states are post-hoc connected to general behaviours (e.g. resting, foraging or travelling). Accelerometers are a common device that is used to record movements of an animal along three axes at very fine temporal scales (e.g. multiple times per second). These devices can collect data over multiple days, often resulting in millions of data points.

In this case study for illustrating spline-based nonparametric HMMs, we consider acceleration data collected for an oceanic whitetip shark at a rate of 16 Hz over 4 days. In order to obtain a single representative metric that incorporates all axes of movement, we used a low-pass filter to remove the static contribution due to gravity, and combined acceleration measurements from the three axes to calculate Overall Dynamic Body Acceleration (ODBA). We further averaged ODBA values over 3s (non-overlapping) windows, resulting in 98,201 observations. ODBA is frequently used as a measure of energy expenditure but we use it as a high-resolution proxy for activity. As these sharks never stop swimming, there is no typical rest state (i.e. ODBA can never be zero) but there are peaks in the time series that indicate bursts in activity.

We developed an HMM with $N = 3$ states to analyse the time series of ODBA values, with the states reflecting general levels of activity. Due to some extreme values in the data, we first log

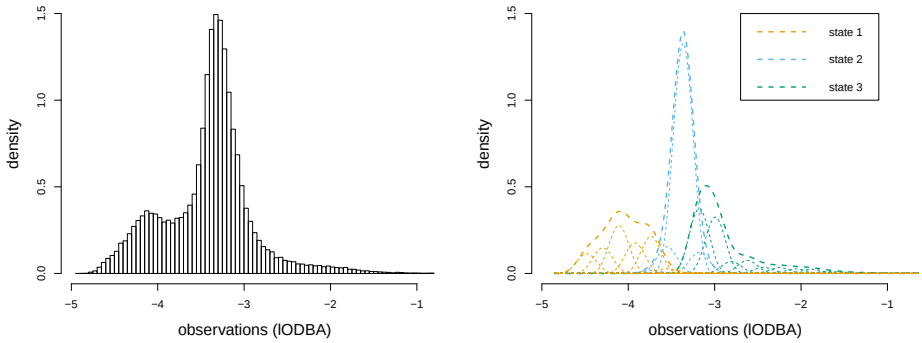

FIGURE 4 Left plot: Histogram of lODBA values. Right plot: 3-state nonparametric HMM fitted to the time series of lODBA values; the state-dependent densities, and also the associated weighted B-spline basis functions used to build these densities, here are weighted by the corresponding entries of the stationary distribution under the fitted model.

transformed the ODBA values (lODBA) before fitting the HMM. A visual inspection of the histogram, which is displayed in Figure 4, reveals two clear modes in lODBA values and a long right tail. The first mode reflects the shark exerting less energy and slowly cruising, the second mode reflects more active behaviour, and the values that lie in the right tail reflect the highest energetic movements, i.e. the largest amounts of activity. The three types of activity can also easily be seen in Figure 5, which displays a subset of the time series analysed here. Because of the size of the data set, there is rich information in the data on essentially every single aspect of the shape of the marginal distribution of lODBA values, and it is clear that the features displayed in the histogram cannot be captured well with relatively few (simple) parametric densities. That is, tying simple parametric state-dependent densities directly to biological behaviours is not feasible here. Moreover, attempting to do so could negatively affect the interpretation of the results, and also any potential inference on the state-switching dynamics.

Inference for the 3-state HMM was conducted via direct maximisation of the penalised log-likelihood, conditional on the state-specific smoothing parameters. The smoothing parameters were selected via AIC from a grid of possible values, as described in Langrock *et al.* (2015b). The fitted HMM indicates high persistence in the states — as is typically the case — with the diagonal entries of the t.p.m. estimated as $\hat{\gamma}_{11} = 0.959$, $\hat{\gamma}_{22} = 0.962$ and $\hat{\gamma}_{33} = 0.962$, respectively. The associated stationary distribution of the Markov chain is $\hat{\boldsymbol{\delta}} = (0.45, 0.28, 0.27)$.

With this case study, we demonstrate the feasibility of the nonparametric estimation approach when dealing with (large) real data sets. When combined with diving data (depth collected at 1 Hz; not shown here) it is clear that the shark was in state 1 (low activity) during descent phases of the dive, but in state 2 or 3 during the ascent. This is biologically realistic as sharks are negatively buoyant and need minimal swimming activity during the dive. The analysis also shows that there may be

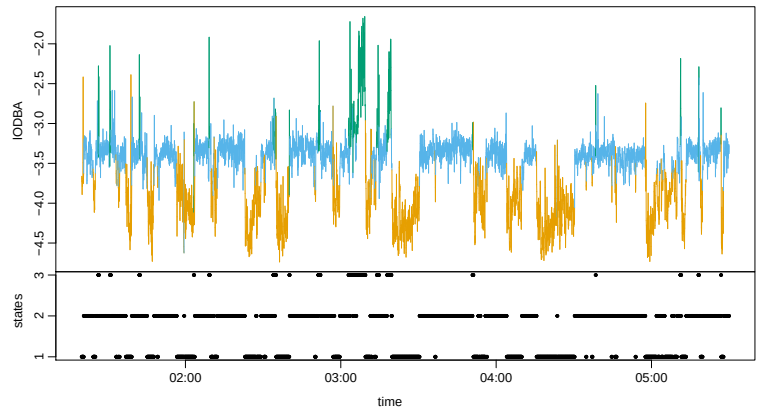

FIGURE 5 Globally decoded time series of lODBA values observed on May 9th, 2014, during the morning hours.

subtle differences in behaviour during the ascent; powered swimming is always needed but some dive ascents include a burst in activity. These could be related to foraging or some other unknown behaviour.

Using nonparametric inference via P-splines for the HMM, we were able to capture the slight multimodality of state 1 and hence more accurately capture the corresponding biological behaviour. We would not have been able to do so within a single state if we had used say a normal distribution. Such a potential lack of flexibility of (parametric) state-dependent distributions can indeed constitute a major problem when making inference related to the state process, ranging from overestimation of the number of states and inaccurate state decoding to invalid inference on covariate influence (Pohle *et al.*, 2017).

3 | STATE-SWITCHING REGRESSION MODELS

3.1 | Markov-switching linear regression models

Markov-switching regression models are closely related to HMMs, but address a slightly different situation. In its most basic (linear) form, a Markov-switching regression model is given as follows:

$$Y_t = \beta_0^{(s_t)} + \beta_1^{(s_t)} x_{t1} + \ldots + \beta_P^{(s_t)} x_{tP} + \sigma^{(s_t)} \epsilon_t, \tag{6}$$

where $\epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ (cf. Hamilton, 1989; Frühwirth-Schnatter, 2006; Kim *et al.*, 2008).

Just like HMMs, a Markov-switching regression model involves a time series $\{Y_t\}_{t=1,\ldots,T}$ and an underlying state process $\{S_t\}_{t=1,\ldots,T}$, but additionally also an associated sequence of covariate vectors $\mathbf{x}_1, \ldots, \mathbf{x}_T$, with $\mathbf{x}_t = (x_{t1}, \ldots, x_{tP})$, affecting the observations. And in contrast to HMMs,

it is not the *density* of the observations Y_t that is (directly) chosen by the underlying state s_t, but instead the *regression function* specifying the effect of $\mathbf{x}_t$ on Y_t, given state s_t. In other words, the linear relationship between $\mathbf{x}_t$ and the mean of Y_t, as well as the residual variance, depends on the state of the underlying hidden Markov chain. The dependence structure in Markov-switching regression models is illustrated in Figure 6.

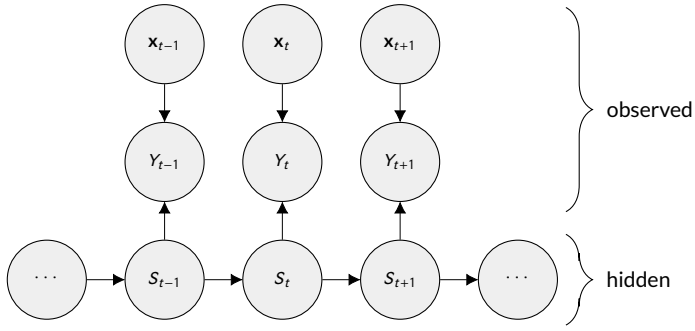

FIGURE 6 Dependence structure in Markov-switching regression models.

Markov-switching regression models are particularly popular in economics, where, for example, during a recession the effect of some explanatory variables $\mathbf{x}_t$ on some economic indicator Y_t might be very different compared to times of economic growth (Hamilton, 2008). The states are usually persistent in the sense that the corresponding *regimes* tend to be active for much longer periods of time than they would be if there was no serial correlation in the mechanism selecting the regimes.

3.2 | Motivation for nonparametric inference within Markov-switching regression models

The motivation for considering nonparametric estimation within Markov-switching regression models is essentially analogous to that given in case of nonparametric HMMs (Section 2.2). For scenarios in which Markov-switching regression models shall be used, one cannot consider scatter plots of the $(\mathbf{x}_t, y_t)$ tuples separately for each state before fitting a model, since it is unknown which observations are associated with which underlying state. This renders it difficult to choose *a priori* if a simple linear state-dependent predictor will be sufficient. In practical applications of Markov-switching regression models, linear predictors do however tend to be used with little or no investigation into their suitability.

Figure 7 illustrates how difficult it can be to formulate an adequate Markov-switching regression model based on exploratory data analysis. The left plot displays a scatter plot of 1000 pairs of observations (x_t, y_t), $t = 1, \ldots, 1000$, generated from a 2-state Markov-switching regression model. Based on this scatter plot, it is next to impossible to decide whether or not linear state-dependent predictors may be adequate. (In fact, it is not even clear at all from this plot alone why a Markov-

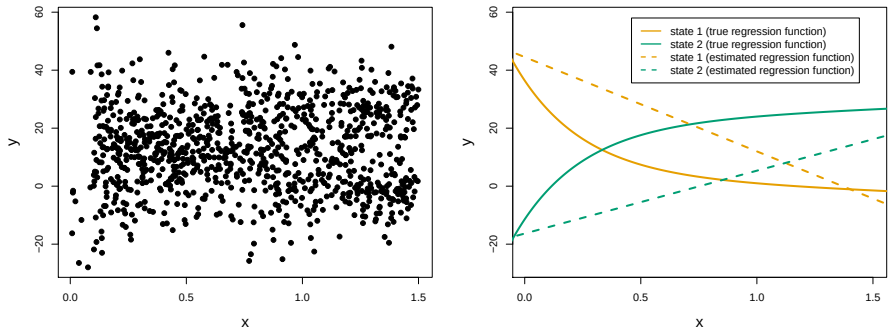

FIGURE 7 Left plot: scatter plot of 1000 pairs of observations (x_t, y_t) generated from a 2-state Markov-switching regression model. Right plot: true state-dependent regression functions used to generate the data (solid lines) and estimated linear state-dependent regression functions (dashed lines).

switching regression model could be a suitable model — this insight could be gleaned from additional inspections of the sample autocorrelation function, or perhaps could have been drawn from expert knowledge.) The actual state-dependent regression functions used to generate the data here were highly nonlinear in both states; see the right plot in Figure 7. Figure 7 also displays the consequences of fitting a misspecified 2-state Markov-switching *linear* regression model, as given in (6) (case $P = 1$), to these data. Here the deviation of the fitted from the actual state-dependent regression functions is substantial, such that it is obvious that the possible consequences listed in Section 2.2 for the case of HMMs — poor predictive performance, frequent state misclassification, invalid inference on state-switching dynamics and on N — would clearly arise also in this scenario.

For Markov-switching regression models, the most obvious way to overcome insufficient flexibility of linear state-dependent predictors is to use polynomial predictors instead. This in many cases will indeed be the best strategy, and we recommend this to be explored before considering much more complicated nonparametric estimation. As in standard regression scenarios (without time series structure and regime shifts), the estimation of polynomial regression functions quickly becomes unstable near the boundaries of the support, and generally highly sensitive to outliers, when increasing the degree of the polynomial. In addition, especially with many covariates in the model it can be cumbersome to explore all plausible models, with different polynomial degrees. Thus, it may sometimes be preferable to directly resort to a nonparametric approach instead, as discussed in the subsequent section.

3.3 | B-spline-based nonparametric model formulation

We initially present our nonparametric estimation approach for the case of normally distributed errors, thus replacing the linear predictor in (6); in Sections 3.5 and 3.6, this model formulation will be extended to allow for other response distributions. For now, we consider the following flexible extension of (6), replacing the (state-dependent) linear effects of all covariates by (state-dependent) smooth effects:

$$Y_t = \underbrace{\beta_0^{(s_t)} + f_1^{(s_t)}(x_{t1}) + \underbrace{f_2^{(s_t)}(x_{t2}) + \ldots + f_P^{(s_t)}(x_{tP})}_{=\eta^{(s_t)}(\mathbf{x}_t)} + \sigma^{(s_t)}\epsilon_t}_{},$$

where $\beta_0^{(i)}$, $i = 1, \ldots, N$, are state-dependent intercepts, and where $\eta^{(s_t)}(\mathbf{x}_t)$ is a shorthand notation for the state-dependent predictor. An implicit assumption made here is that the state-dependent effects of the covariates $x_{t1}, \ldots, x_{tP}$ are *additive* (in the sense of generalized additive models, see Wood, 2017). For flexible estimation, we express each of the functions $f_p^{(i)}$, $i = 1, \ldots, N$, $p = 1, \ldots, P$, as a finite linear combination of basis functions, $\phi_1, \ldots, \phi_K$:

$$f_p^{(i)}(x) = \sum_{k=1}^{K} \omega_{i,p,k}\phi_k(x).$$

For each of the functions $f_p^{(i)}$, we need to fix one of the coefficients in order to render the model identifiable.

For the same reasons as in Section 2.3, the examples below use cubic B-splines as basis functions $\phi_1, \ldots, \phi_K$. The B-spline basis size, K, determines the flexibility of the functional form. However, instead of trying to select an optimal K, we again follow the spline literature (e.g., Eilers and Marx, 1996, Wood, 2017) and penalise the likelihood, thus the basis size simply needs to be sufficiently large in order to ensure sufficient flexibility.

Using the toy example above (Figure 7), with only one covariate and $N = 2$ states, Figure 8 illustrates how the state-dependent smooth regression functions are built as linear combinations of weighted B-splines. Fitting this model to the simulated data was again achieved by numerically maximising the penalised log-likelihood. More details on inference in these classes of models is provided in the subsequent section.

3.4 | Inference

3.4.1 | Direct numerical maximum penalised likelihood

Likelihood evaluation in Markov-switching regression models, either parametric or nonparametric, is performed using the forward algorithm, completely analogous as in case of HMMs. In particular, the forward recursion, and ultimately the calculation of the likelihood of a Markov-switching regression

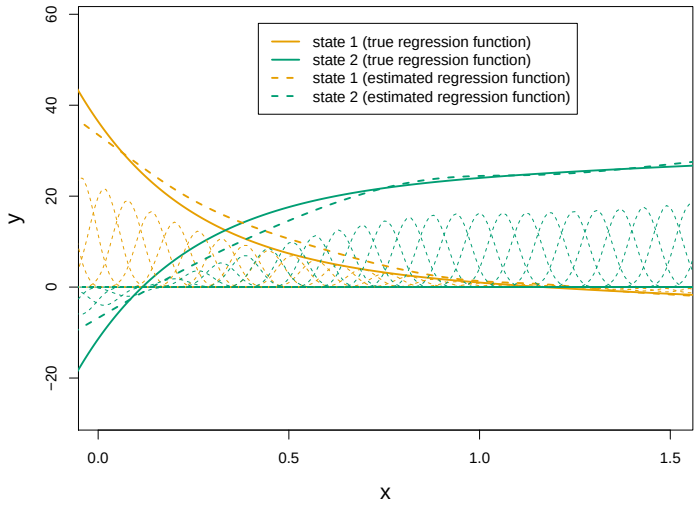

FIGURE 8 True state-dependent regression functions (solid lines) and state-dependent regression functions estimated using linear combinations of weighted B-splines (dashed lines). The thin dashed lines display the contributions of the individual B-spline basis functions to the estimators of the regression functions.

model, proceed exactly as in (2) and (3), respectively, where now

$$\mathbf{P}(y_t) = \text{diag}\big(f_{\mathcal{N}\left(\eta^{(1)}(\mathbf{x}_t),\sigma^{(1)}\right)}(y_t), \ldots, f_{\mathcal{N}\left(\eta^{(N)}(\mathbf{x}_t),\sigma^{(N)}\right)}(y_t)\big),$$

using the notation $f_{\mathcal{N}(\mu,\sigma)}$ to denote the density of a normal distribution with mean μ and standard deviation σ.

Again analogously as in case of nonparametric HMMs, the log-likelihood of the nonparametric Markov-switching regression model is modified by including a difference penalty, one for each smooth function to be estimated:

$$l_{\text{pen}} = \log \mathcal{L} - \sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\lambda_{ip}}{2}\sum_{k=3}^{K}(\Delta^2\omega_{i,p,k})^2 \,.$$

Second-order differences are appealing here since the basic linear Markov-switching regression model is then recovered as $\lambda_{ip} \to \infty \; \forall i, p$, as the penalty then suppresses any nonlinear effects of the $\omega_{i,p,k}$.

Generally, the penalised maximum likelihood estimate obtained will reflect a compromise be-

tween goodness of fit and smoothness, respectively, with the smoothing parameters λ_{ip}, $i = 1, \ldots, N$, $p = 1, \ldots, P$, controlling the wiggliness of the smooth terms. Smoothing parameter selection can be conducted using cross-validation techniques or model selection criteria (see Section 2.4.1).

3.4.2 | Estimation using the EM algorithm

Alternatively, maximum penalised likelihood estimates for state-switching regression models can be obtained via the EM algorithm. As in the case of state-switching density models, the idea is to maximise the CDPLL. The likelihood structure is identical to that in case of state-switching density models, such that the likelihood differs only in the specific form of state-dependent distributions, $f_i(y_t)$:

$$l_{c,\text{pen}} = \sum_{i=1}^{N} u_i(1) \log(\delta_i) + \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{t=2}^{T} v_{ij}(t) \log(\gamma_{ij}) + \sum_{i=1}^{N} \sum_{t=1}^{T} u_i(t) \log \left(f_{\mathcal{N}\left(\eta^{(i)}(\mathbf{x}_t), \sigma^{(i)}\right)}(y_t) \right) - \sum_{i=1}^{N} \sum_{p=1}^{P} \frac{\lambda_{ip}}{2} \sum_{k=3}^{K} (\Delta^2 \omega_{i,p,k})^2.$$

The E step again involves calculating the conditional expectations of the $u_i(t)$ and $v_{ij}(t)$, respectively, given the data and the current parameter estimates, which is completely analogous as described in Section 2.4.2.

The M step involves the maximization of the CDPLL with respect to the model parameters, the first two steps of which — i.e. the maximisation with respect to δ_i and γ_{ij}, $i, j = 1, \ldots, N$ — proceed exactly as described in Section 2.4.2. The third step, i.e. the maximisation with respect to the $\omega_{i,p,k}$, reduces to a standard regression problem with weighted observations, since the last two terms in the CDPLL (the only terms which depend on the $\omega_{i,p,k}$) are equivalent to the weighted penalised log-likelihood of a standard regression model, where (for each state i) the contribution of the t-th observation to the log-likelihood is weighted by the current value of $u_i(t)$. For given values of the λ_{ip}, this part of the M step can therefore be conducted using penalised weighted least squares estimation. To avoid a computationally intensive grid search based on generalised cross-validation or information criteria for selecting the λ_{ip}, we propose to estimate the λ_{ip} *within* each M step, using the linear mixed model representation of P-splines, proceeding in a similar way as outlined in Section 2.4.2 (see e.g. Equation (13) in Kauermann, 2005, for details).

3.5 | Extension to Markov-switching generalised additive models

It is completely straightforward, both in terms of the model formulation and also regarding the associated inferential techniques and implementation, to extend Markov-switching regression models as discussed in Sections 3.1 (parametric case) and 3.3 (nonparametric case) to models with response distributions other than the normal. In the nonparametric modelling framework, we assume that, conditional on s_t and $\mathbf{x}_{t\cdot}$, Y_t follows some distribution, e.g. from the exponential family, and specify

the model for the mean as

$$g\left(\mathbb{E}(Y_t)\right) = \eta^{(s_t)}(\mathbf{x}_t) = \beta_0^{(s_t)} + f_1^{(s_t)}(x_{t1}) + f_2^{(s_t)}(x_{t2}) + \ldots + f_P^{(s_t)}(x_{tP}),$$

where g is some link function associated with the distribution considered, and where the smooth functions $f_p^{(i)}$, $i = 1, \ldots, N$, $p = 1, \ldots, P$, are modelled as linear combinations of a large number of B-splines, with the associated weights being estimated. For distributions with additional dispersion parameters, such as the normal or the gamma distribution, those are also modelled as dependent on the state s_t. Likelihood evaluation and inference proceeds analogously as detailed in Section 3.4, with the obvious changes to $\mathbf{P}(y_t)$ (see Langrock *et al.*, 2017, for details). If the EM algorithm is used for model fitting, efficient software which exploits the linear mixed model representation of P-splines can be employed to estimate the λ_{ip} and $\omega_{i,p,k}$ in the M step, e.g. the gam function from the mgcv package (Wood, 2017).

3.6 | Extension to Markov-switching generalised additive models for location, scale and shape

The models from the previous section can be further extended by also modelling state-dependent parameters of the response distribution beyond the mean, e.g. variance, skewness and kurtosis parameters, but also zero inflation and other parameters, as smooth functions of a given set of explanatory variables (Adam *et al.*, 2017). Corresponding Markov-switching generalised additive models for location, scale and shape (MS-GAMLSS) can for example account for heteroscedasticity even within states, as the variance parameter is assumed to be some function of the covariates rather than being constant.

In an MS-GAMLSS we assume that, conditionally on s_t and $\mathbf{x}_t$, Y_t follows some parametric distribution $f_{s_t}(y_t, \mu_t^{(s_t)}, \sigma_t^{(s_t)}, \nu_t^{(s_t)}, \tau_t^{(s_t)})$ (which does not necessarily need to belong to the exponential family), and specify a model for each distribution parameter:

$$\begin{aligned}
g_\mu(\mu_t^{(s_t)}) &= \eta_\mu^{(s_t)}(\mathbf{x}_t) = \beta_{\mu 0}^{(s_t)} + f_{\mu 1}^{(s_t)}(x_{t1}) + f_{\mu 2}^{(s_t)}(x_{t2}) + \cdots + f_{\mu P}^{(s_t)}(x_{tP});\\
g_\sigma(\sigma_t^{(s_t)}) &= \eta_\sigma^{(s_t)}(\mathbf{x}_t) = \beta_{\sigma 0}^{(s_t)} + f_{\sigma 1}^{(s_t)}(x_{t1}) + f_{\sigma 2}^{(s_t)}(x_{t2}) + \cdots + f_{\sigma P}^{(s_t)}(x_{tP});\\
g_\nu(\nu_t^{(s_t)}) &= \eta_\nu^{(s_t)}(\mathbf{x}_t) = \beta_{\nu 0}^{(s_t)} + f_{\nu 1}^{(s_t)}(x_{t1}) + f_{\nu 2}^{(s_t)}(x_{t2}) + \cdots + f_{\nu P}^{(s_t)}(x_{tP});\\
g_\tau(\tau_t^{(s_t)}) &= \eta_\tau^{(s_t)}(\mathbf{x}_t) = \beta_{\tau 0}^{(s_t)} + f_{\tau 1}^{(s_t)}(x_{t1}) + f_{\tau 2}^{(s_t)}(x_{t2}) + \cdots + f_{\tau P}^{(s_t)}(x_{tP}).
\end{aligned}$$

Likelihood evaluation and inference proceeds completely analogously as detailed in Section 3.4, with the obvious changes to $\mathbf{P}(y_t)$ and the penalty term. In particular, we now require separate smoothing parameters for each parameter of the distribution considered. Again, when using the EM algorithm, we can make use of available software for estimating the λ_{ip} and $\omega_{i,p,k}$ in the M step, e.g. the gamlss function from the package of the same name (Stasinopoulos *et al.*, 2017).

3.7 | Case study: Spanish energy prices

To illustrate the application of Markov-switching regression models, we model the relationship between the daily average price of energy in Spain, y_t, and the oil price, x_t, over time. The data cover all 1,761 working days between February 1, 2002, and October 31, 2008 and are available in the R package MSwM (Sanchez-Espigares and Lopez-Moreno, 2014). Markov-switching regression models play an important role in modelling financial time series data in general, and energy market data in particular, as observations are typically driven by market characteristics (states) that are not directly observable (Huisman and Mahieu, 2003; Eichler and Tuerk, 2013). In our example, for instance, it seems plausible that the relationship between energy and oil price may vary across different market regimes, e.g. recessions as opposed to periods of economic growth. Accounting for such regime shifts is important for forecasts, as neglecting these features in the model formulation may lead to an over- or underestimation of the energy prices.

To demonstrate potential advantages of a B-spline-based nonparametric model formulation, we consider three different models. As a benchmark model, we fitted a 2-state Markov-switching *linear* model (MS-LIN) with state-dependent linear predictor $\eta^{(s_t)}(x_t) = \beta_0^{(s_t)} + \beta_1^{(s_t)} x_t$ for the conditional mean and state-dependent (constant) variance $\sigma^{(s_t)}$. Additionally, we fitted a 2-state MS-GAM with state-dependent *nonlinear* predictor $\eta^{(s_t)}(x_t) = \beta_0^{(s_t)} + f_1^{(s_t)}(x_t)$ for the conditional mean and (constant) state-dependent variance $\sigma^{(s_t)}$, assuming a normal distribution for the energy prices. Finally, we fitted a 2-state MS-GAMLSS, specifying state-dependent *nonlinear* predictors of the form $\eta^{(s_t)}(x_t) = \beta_0^{(s_t)} + f_1^{(s_t)}(x_t)$ for *both* the conditional mean, $\mu_t^{(s_t)}$, *and* the conditional variance, $\sigma_t^{(s_t)}$, hence allowing not only the mean but also the variance to depend on x_t. All models were fitted using the EM algorithm, as described in Sections 3.4.2, 3.5 and 3.6.

The three models fitted to the data are illustrated in Figure 9. The left panels display the state-dependent predictors for the conditional mean, and additionally several quantiles of the state-dependent distributions of the response Y_t, under the fitted models. The right panels display the associated (locally) decoded time series of energy prices, indicating the most likely states at any time. The diagonal entries of the t.p.m. were estimated as $\hat{\gamma}_{11} = 0.991$ and $\hat{\gamma}_{22} = 0.990$ (MS-LIN), $\hat{\gamma}_{11} = 0.991$ and $\hat{\gamma}_{22} = 0.991$ (MS-GAM) and $\hat{\gamma}_{11} = 0.994$ and $\hat{\gamma}_{22} = 0.991$ (MS-GAMLSS), respectively, indicating strong persistence within the states and only minor differences in the state processes of the three models fitted.

The MS-LIN substantially overestimates the conditional mean particularly for high values of the covariate x_t, primarily due to the lack of flexibility of the state-dependent predictor. Furthermore, both the MS-LIN and the MS-GAM substantially underestimate the variance for oil prices between 40 and 60 USD, where a considerable proportion of the observations lie above (below) the 0.95 (0.05) quantile of the fitted state-dependent distributions, while overestimating the variance for oil prices greater than 70 USD, where almost all observations lie below (above) the 0.95 (0.05) quantile. Thus, although the MS-GAM accurately captures the conditional mean, the overall distribution is not completely modelled. This can be problematic if the interest lies not only in the conditional mean but also in the conditional quantiles of the fitted state-dependent distributions, which is often the case when modelling financial data. By modelling not only the conditional mean but also the conditional

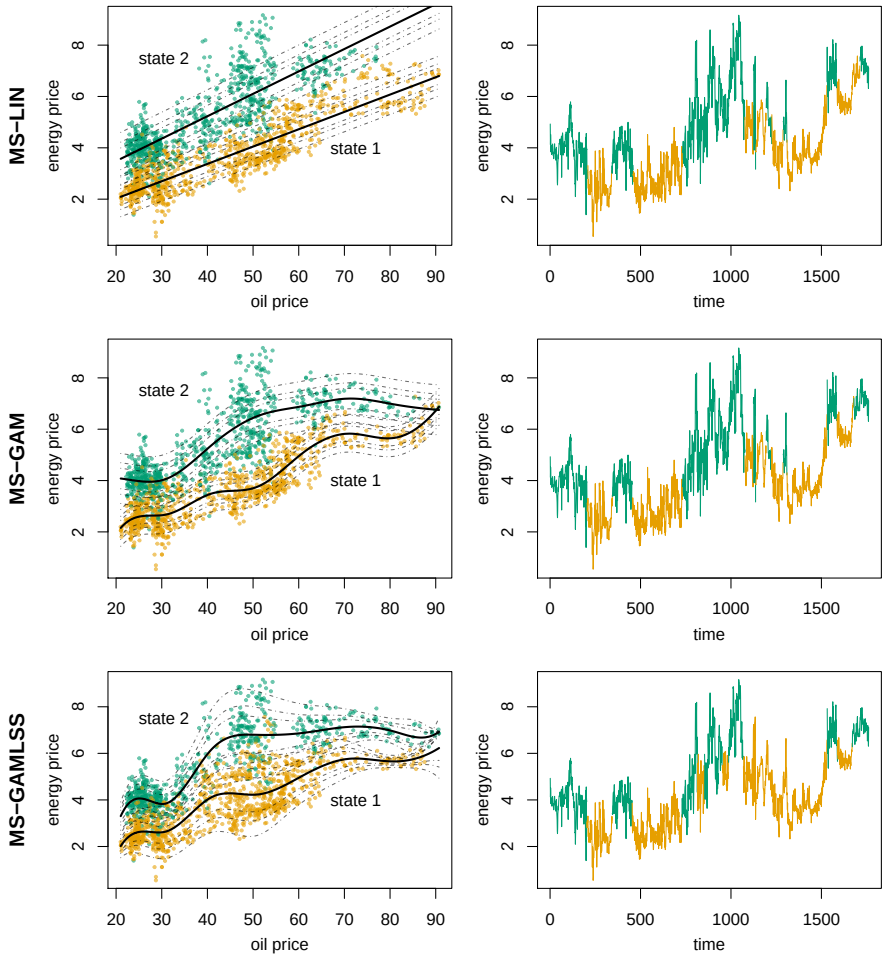

FIGURE 9 Plots on the left: estimated state-dependent predictors for the conditional mean (solid lines) along with the 0.05, 0.15, 0.25, 0.75, 0.85 and 0.95 quantiles of the fitted state-dependent distributions (dashed lines) and the (locally) decoded observations. In the case of the MS-GAMLSS, the respective quantiles were computed using the estimated state-dependent predictors for the conditional variance. Plots on the right: the corresponding (locally) decoded time series of energy prices. Orange points and lines correspond to state 1, while green points and lines correspond to state 2.

variance as functions of x_t, the MS-GAMLSS is able to overcome these caveats of the simpler (nested) models MS-LIN and MS-GAM.

At this point we ought to stress that while the goodness of fit of the MS-GAMLSS here is clearly superior to that of the alternative models, there is undoubtedly some overfitting involved, and it is

not at all clear if the MS-GAMLSS in the given setting would perform best out of sample. For the given time series, where we see only about 11-12 switches between the two regimes, a model as flexible as the MS-GAMLSS is almost certainly overparameterised, i.e. too complex. Nevertheless, the case study clearly demonstrates the appeal of being able to conduct such flexible modelling, even if in practice the full flexibility of these classes of models should probably be exploited only if larger data sets are available.

4 | DISCUSSION

Nonparametric inference — via P-splines as discussed in this paper or using other techniques, such as kernels (see Piccardi and Perez, 2007) or smoothed histograms (as suggested in Eilers and Marx, 1996) — adds substantial modelling flexibility to state-switching density and regression models. Corresponding models effectively enlarge the class of state-switching models, offering versatile new tools for statistical modelling of time series data. Viewed from a different angle, these models extend well-established nonparametric density and regression models to scenarios with time series structure, where regime shifts drive the serial correlation.

The increased flexibility does of course come at a cost: the necessity to select smoothing parameters leads to a notable increase in the computational effort, and in addition the nonparametric model formulations are numerically less stable than their parametric counterparts. Thus, while clearly beneficial in some scenarios, one should always carefully consider if parametric models really are insufficient before working with the more challenging models discussed in this paper.

Regarding the selection of smoothing parameters, there has been considerable effort spent in the GAM literature on developing efficient schemes for estimating these; examples include Wood (2004), Wood (2008), Wood (2011) and Wood *et al.* (2016). Adapting these procedures for the state-switching modelling frameworks considered here might prove fruitful. Taking a mixed effects view of the model (as seen above for the EM fitting routine), where the penalties can be viewed as prior precision matrices — potentially improper, though this can be remedied, see, e.g., Marra and Wood (2011) — allows for fitting and inference using general-purpose modelling software such as Stan (Carpenter *et al.*, 2017) and Template Model Builder (Kristensen *et al.*, 2016).

ACKNOWLEDGEMENTS

The authors are grateful to Marco Grzegorczyk for organising this special issue, and to Paul Eilers and Thomas Kneib for providing some very helpful comments.

REFERENCES

Adam, T., Mayr, A. & Kneib, T. (2017), Gradient boosting in Markov-switching generalized additive models for location, scale and shape. *arXiv*, 1710.02385.

Alexandrovich, G., Holzmann, H. & Leister, A. (2016) Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, **103**, 423–434.

Altman, R.M. (2007), Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**, 201–210.

Baum, L.E., Petrie, T., Soules, G. & Weiss, N. (1970), A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Bebbington, M.S. (2007), Identifying volcanic regimes using hidden Markov models. *Geophysical Journal International*, **171**, 921–942.

Borchers, D.L., Zucchini, W., Heide-Jørgensen, M.P., Cañadas, A. & Langrock, R. (2013), Using hidden Markov models to deal with availability bias on line transect surveys. *Biometrics*, **69**, 703–713.

Bulla, J. & Bulla, I. (2006), Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics & Data Analysis*, **51**, 2192–2209.

Bulla, J. & Berzel, A. (2008), Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, **23**, 1–18.

Bulla, J., Lagona, F., Maruotti, A. & Picone, M. (2012), A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 544–567.

de Boor, C. (1978), A Practical Guide to Splines. Springer, Berlin.

Carpenter B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & and Riddell, A. (2017) Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977), Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Eichler, M. & Tuerk, D. (2013), Fitting semiparametric Markov regime-switching models to electricity spot prices. *Energy Economics*, **36**, 614–624.

Eilers, P.H.C. & Marx, B.D. (1996), Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**, 89–121.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York, NY: Springer.

Green, P.J. & Silverman, B.W. (1994), Nonparametric regression and generalized linear models: a roughness penalty approach. Chapman & Hall.

Guédon, Y. (2003), Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, **12**, 604–639.

Hamilton, J.D. (1989), A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357—384.

Hamilton, J.D. (2008), Regime-switching models, In: *The New Palgrave Dictionary of Economics*, Durlauf, S.N. and Blume, L.E. (eds.), Second Edition.

Huisman, R. & Mahieu, R. (2003), Regime jumps in electricity prices. *Energy Economics*, **25**, 425–434.

Kim, C.-J., Piger, J. & Startz, R. (2008), Estimation of Markov regime-switching regression models with endogenous switching, *Journal of Econometrics*, **143**, 263–273.

Holzmann, H. & Schwaiger, F. (2015), Hidden Markov models with state-dependent mixtures: minimal representation, model testing and applications to clustering. *Statistics and Computing*, **25**, 1185–1200.

Juang, B.H. & Rabiner, L.R. (1991), Hidden Markov models for speech recognition. *Technometrics*, **33**, 251–272.

Kauermann, G. (2005), A note on smoothing parameter selection for penalized spline smoothing. *Journal of statistical planning and inference*, **127**, 53–69.

Kock, A., O'Riain, M.J., Mauff, K., Meÿer, M., Kotze, D. & Griffiths, C. (2013), Residency, habitat use and sexual segregation of white sharks, *Carcharodon carcharias* in False Bay, South Africa. *PLOS ONE*, **8**, e55048.

Kristensen, K., Nielsen, A,. Berg, C.W., Skaug, H. & Bell, B.M. (2016), TMB: Automatic Differentiation and Laplace Approximation. *Journal of Statistical Software*, **70**, 1–21.

Lamb, J.S., Satgé, Y.G. & Jodice, P.G.R. (2017), Influence of density-dependent competition on foraging and migratory behavior of a subtropical colonial seabird. *Ecology and Evolution*, **7**, 6469–6481.

Langrock, R., Swihart, B.J., Caffo, B.S., Crainiceanu, C.M. & Punjabi, N.M. (2013), Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine*, **32**, 3342–3356.

Langrock, R., Michelot, T., Sohn, A. & Kneib, T. (2015a), Semiparametric stochastic volatility modelling using penalized splines. *Computational Statistics*, **30**, 517–537.

Langrock, R., Kneib, T., Sohn, A. & DeRuiter, S.L. (2015b), Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, **71**, 520–528.

Langrock, R., Kneib, T., Glennie, R. & Michelot, T. (2017), Markov-switching generalized additive models. *Statistics and Computing*, **27**, 259–270.

Leos-Barajas, V., Gangloff, E.J., Adam, T., Langrock, R., Morales, J.M., van Beest, F.M. & Nabe-Nielsen, J. (2017a), Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 232–248.

Leos-Barajas, V., Photopoulou, T., Langrock, R., Patterson, T.A., Watanabe, Y., Murgatroyd, M. & Papastamatiou, Y. (2017b), Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, **8**, 161–173.

Marra, G. & Wood, S.N. (2011), Practical Variable Selection for Generalized Additive Models. *Computational Statistics and Data Analysis*, **55**, 2372—2387.

MacDonald, I.D. (2014), Numerical maximisation of the likelihood: a neglected alternative to EM?. *International Statistical Review*, **82**, 296–308.

Maruotti, A. (2011), Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review*, **79**, 427–454.

Michelot, T., Langrock, R. & Patterson, T.A. (2016), moveHMM: An R package for analysing animal movement data using hidden Markov models. *Methods in Ecology and Evolution*, **7**, 1308–1315.

Morellet, N., Bonenfant, C., Börger, L., Ossi, F., Cagnacci, F., Heurich, M., Kjellander, P., Linnell, J.D.C., Nicoloso, S., Sustr, P., Urbano, F. & Mysterud, A. (2013), Seasonality, weather and climate affect home range size in roe deer across a wide latitudinal gradient within Europe. *Journal of Animal Ecology*, **82**, 1326–1339.

Piccardi, M. & Perez, O. (2007), Hidden Markov models with kernel density estimation of emission probabilities and their use in activity recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Pohle, J., Langrock, R., van Beest, F.M. & Schmidt, N.M. (2017), Selecting the number of states in hidden Markov models — pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 270–293.

Sanchez-Espigares, J.A. & Lopez-Moreno, A. (2014), MSwM: Fitting Markov-Switching Models. R package version 1.2. http://CRAN.R-project.org/package=MSwM.

Schliehe-Diecks, S., Kappeler, P.M. & Langrock, R. (2012), On the application of mixed hidden Markov models to multiple behavioural time series. *Interface Focus*, **2**, 180–189.

Schellhase, C. & Kauermann, G. (2012), Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, **27**, 757–777.

Sherlock, C., Xifara, T., Telfer, S. & Begon, M. (2013), A coupled hidden Markov model for disease interactions. *Journal of the Royal Statistical Society: Series C*, **62**, 609–627.

Stasinopoulos, M.D., Rigby, R.A., Heller, G.Z., Voudouris, V. & De Bastiani, F. (2017), Flexible Regression and Smooting: Using GAMLSS in R. Chapman & Hall/CRC, Boca Raton.

Visser, I., Raijmakers, M.E.J. & and Molenaar, P.C.M. (2002), Fitting hidden Markov models to psychological data. *Scientific Programming*, **10**, 185–199.

Volant, S., Bérard, C., Martin-Magniette, M.-L. & Robin, S. (2015), Hidden Markov models with mixtures as emission distributions. *Statistics and Computing*, **24**, 493–504.

Welch, L.R. (2003): Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Th. Society Newsletter*, **53**, 10–13.

Wood, S.N. (2004) Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association*, **99**, 673–686.

Wood, S.N. (2008) Fast Stable Direct Fitting and Smoothness Selection for Generalized Additive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**, 495–518.

Wood, S.N. (2011) Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semi-parametric Generalized Linear Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**, 3–36.

Wood, S.N., Pya, N. & Säfken, B. (2016) Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association*, **111**, 1–45.

Wood, S.N. (2017), *Generalized Additive Models: An Introduction with R, Second Edition*, Boca Raton, FL: Chapman & Hall/CRC.

Zucchini, W., MacDonald, I.L. & Langrock, R. (2016), Hidden Markov Models for Time Series: An Introduction using R, 2nd Edition. Chapman & Hall/CRC, Boca Raton.

Zucchini, W., Raubenheimer, D. & MacDonald, I.L. (2008), Modeling time series of animal behavior by means of a latent-state model with feedback. *Biometrics*, **64**, 907–815.